ANALYZE HOUSE SALES IN A NORTHWESTERN COUNTY USING REGRESSION MODEL

Nishmitha Naik

19th May 2024

Summary

- **Objective:**
- Develop a predictive model to estimate the potential increase in home sale prices based on specific renovation features or upgrades.
- **Approach:**
- Leverage historical sales data from King County.
- Apply regression modelling techniques to analyze the impact of various renovations.
- **Goal:**
- Provide homeowners with actionable recommendations on renovations that yield significant ROI in terms of increased property value.

Outline

- Business Problem
- Data
- Results
- Conclusions

Business Problem

Maximizing Property Value through Renovations

- **Real Estate Agency's Aim:**

- Provide valuable insights and guidance to homeowners in King County to maximize their property values.

- **Challenge:**

- Advising homeowners on the potential impact of various home renovations.
- Quantifying the extent to which specific renovations can increase property value

Data & Methods

- The data files provide the foundation for analyzing historical sales data from King County, including information on property features, sale price, and any renovation or upgrade details.
- The above data files contains information's like bedrooms, bathrooms, sqft_living etc.

Regression Modelling

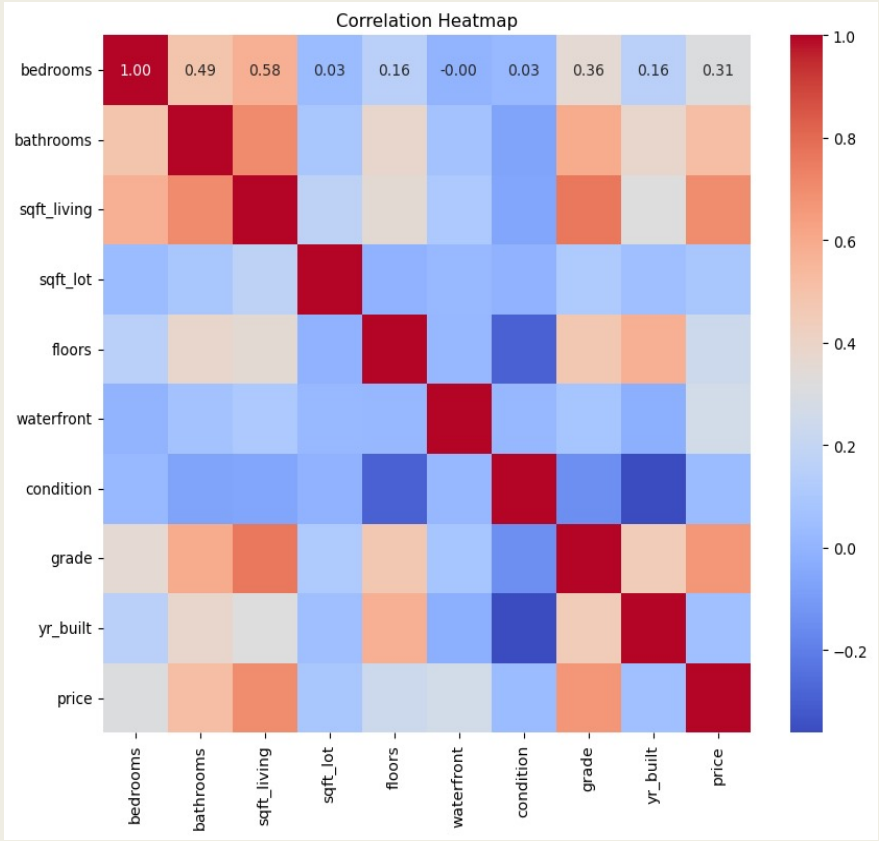
Lets define the target variable Price and select the features for your regression analysis, follow these steps:

1. **Target Variable (Dependent Variable):** The ‘Price' column, as it represents the sale price of the house.
2. **Features (IndependentVariables):**bedrooms, bathrooms, sqft_living, sqft_lot, floor, waterfront, condition, grade, yr_built.

Exploratory Data Analysis

- **Strong Positive Correlations:**

- **Square Footage of Living Space (sqft_living):**
  - Strongest positive correlation with sale price.
  - Larger living areas significantly increase home value.
- **Number of Bathrooms (bathrooms):**
  - Strong positive correlation with sale price.
  - More bathrooms tend to increase home value.
- **Grade:**
  - Strong positive correlation with sale price.
  - Higher grade homes sell for more.



■ Moderate Positive Correlations:

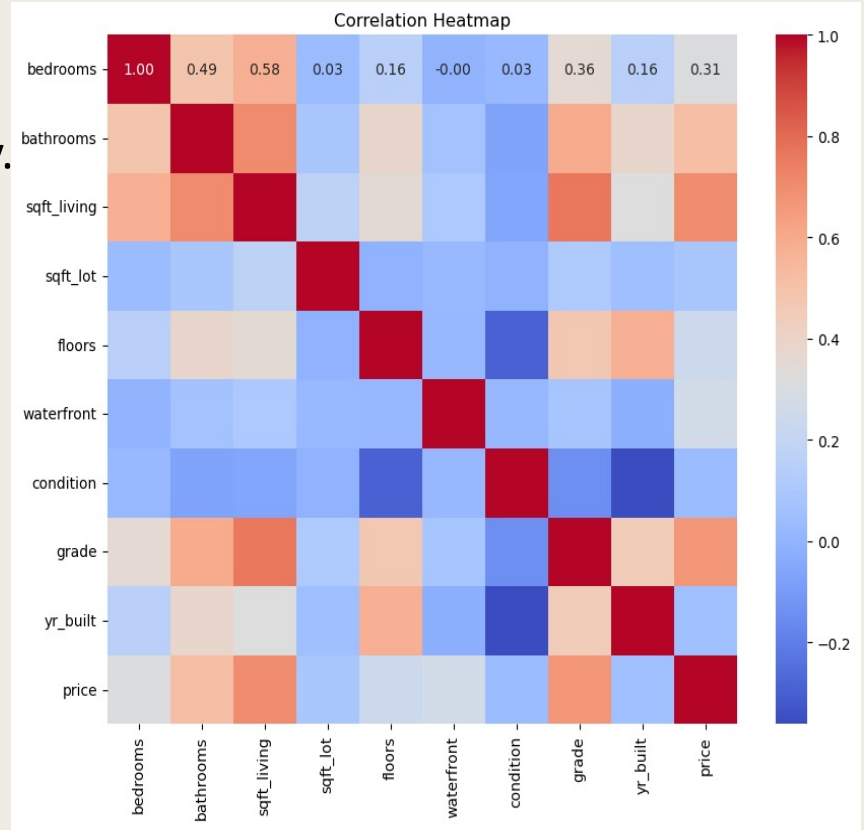
- **Number of Bedrooms (bedrooms):**
  - Moderate positive correlation with sale price.
  - More bedrooms increase home value, but less significantly.
- **Size of Lot (sqft_lot):**
  - Weak positive correlation with sale price.
  - Slight tendency for larger lots to increase home value.

■ Weak Positive Correlations:

- **Number of Floors (floors):**
  - Very weak positive correlation with sale price.
  - Minimal impact on home value.

■ Weak Negative Correlation:

- **Condition of the House (condition):**
  - Weak negative correlation with sale price.
  - Poorer condition slightly decreases home value.



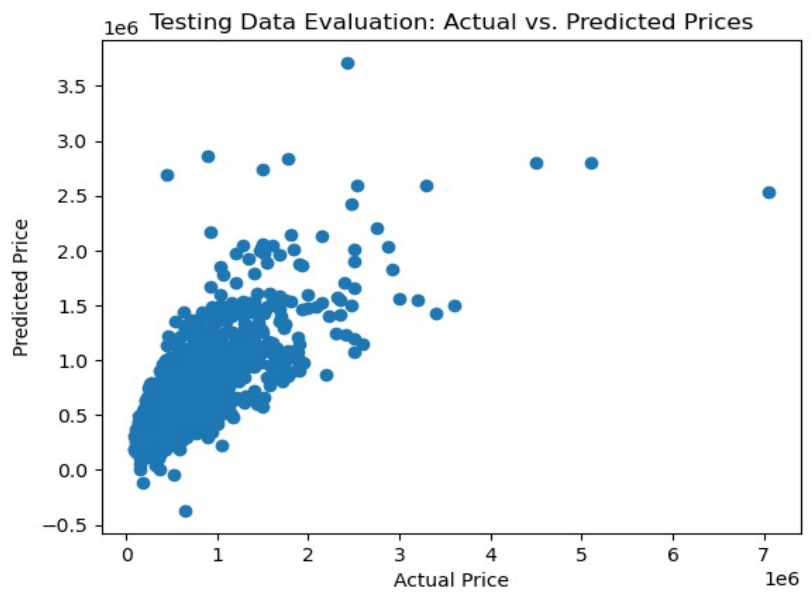
Model Development Process

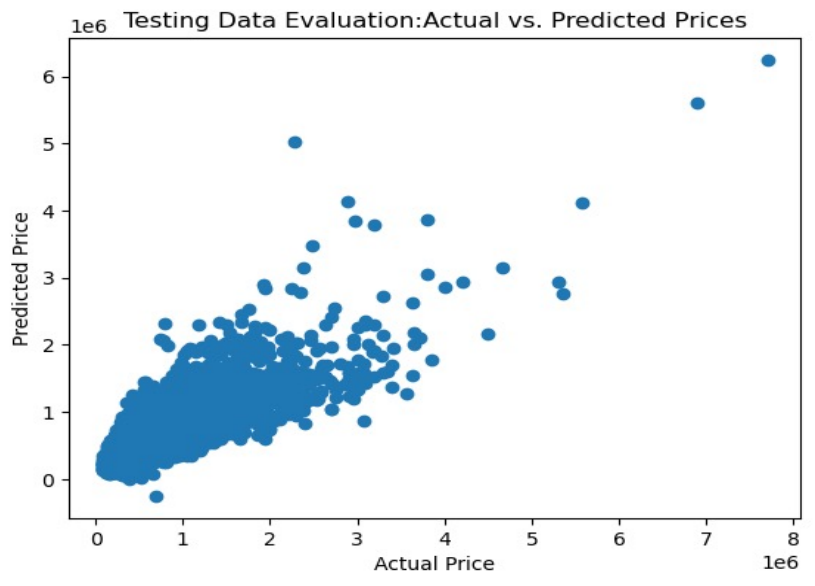
- **Step1 : Train-Test Split**
- **Purpose:**
  - Evaluate model's performance on unseen data.
- **Method:**
  - Use train_test_split() from scikit-learn.
- **Step2 : Linear Regression Model**
- **Initialization:**
  - Use scikit-learn's LinearRegression() class.
- **Step3 : Model Fitting**
- **Training:**
  - Train the model on training data using fit() method.

- **Step4 : Prediction**
- **Testing:**
  - Predict prices on testing data using predict() method.
- **Step5 : Model Evaluation**
- **Metrics:**
  - Calculate MSE, RMSE, and R-squared values.
- **Step6 : Visualization**
- **Scatter Plot:**
  - Visualize relationship between predicted and actual prices.
- **Step7: Interpretation**
- **Feature Coefficients:**
  - Print coefficients to understand feature influence on predictions.

Results

- **Insight 1 :** Testing & Training Data Evaluation





Testing Data Evaluation

- Mean Squared Error: 51056048785.58351
- Root Mean Squared Error: 225955.8558337967
- R-squared: 0.6079134748656662

Training Data Evaluation

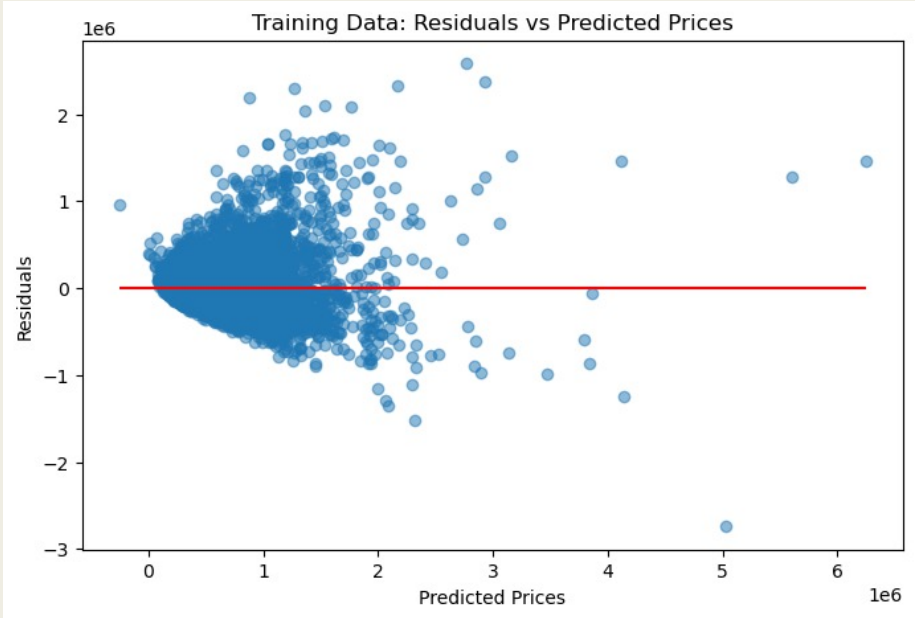
- Mean Squared Error (MSE): 46475275094.99554
- Root Mean Squared Error (RMSE): 215581.24940494137
- R-squared (R^2): 0.6585964821129644

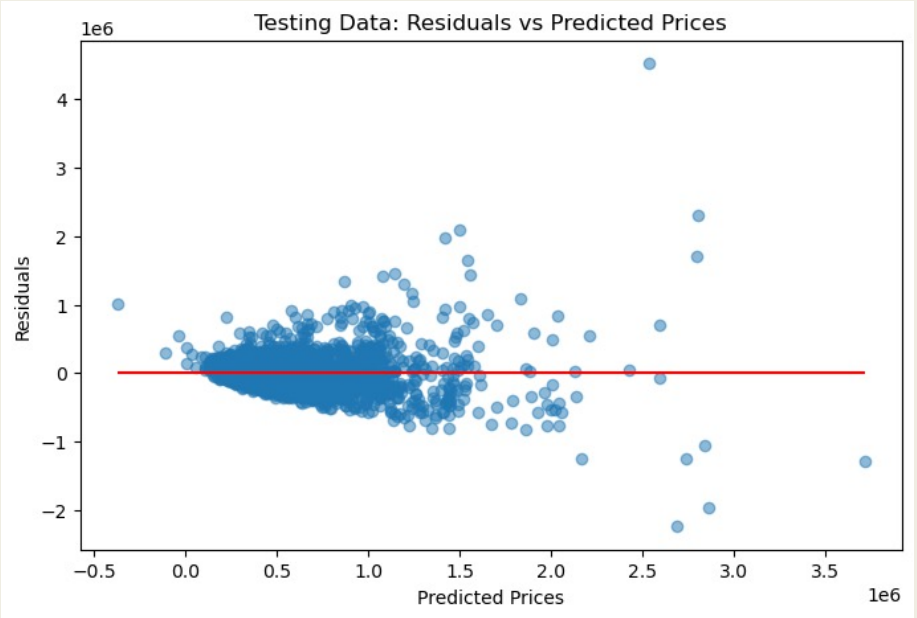
Analysis

- **Model Fit:**
- **R-squared Values:**
  - Indicate a moderate fit for both training and testing datasets.
  - Captures a significant portion of variance in house prices.
  - Room for improvement remains.
- **Generalization:**
- **Performance:**
  - Model performs reasonably well on both training and testing sets.
  - Indicates good generalization to unseen data.
- **Overfitting:**
  - Slightly lower R-squared and higher RMSE for the testing set.
  - Suggests some degree of overfitting.

Results

- **Insight 1 : Testing & Training Data Evaluation**





Training Residuals:
Mean: 7.892011433499979e-10
Standard Deviation:
215587.48864231847

Testing Residuals:
- **Mean:** -5711.565814452754
- **Standard Deviation:**
225909.80614973872

Residual Analysis

- **Training Residuals:**
- **Mean Residual:**
  - Essentially zero, indicating unbiased predictions on training data.
- **Standard Deviation:**
  - 215,587.49, suggesting most predictions deviate from the mean by this amount.
  - Relatively high but contextually dependent on acceptability.
- **Testing Residuals:**
- **Mean Residual:**
  - -5,711.57, indicating the model underestimates house prices by this amount on average.
- **Standard Deviation:**
  - 225,909.81, showing a wider spread of prediction errors compared to training data.

Insights from Regression Modeling Results

- **Mean Squared Error (MSE) and Root Mean Squared Error (RMSE):**
- **High RMSE Values:**
  - Indicate relatively high average prediction error.
  - Useful for estimating average prediction accuracy.
- **R-squared (R^2) Value:**
- **Moderate R-squared:**
  - Explains a significant portion of variability in home prices.
  - Indicates potential for further model improvement.
- **Residual Analysis:**
- **Mean Residuals:**
  - Close to zero for training data, indicating unbiased predictions.
  - Non-zero for testing data, suggesting potential prediction bias.
- **Standard Deviation of Residuals:**
  - High standard deviation reflects variability in home prices.
  - Indicates areas where the model struggles to capture accurate predictions.

Practical Implications for Homeowners Results

- **Focus on Key Areas:**
- **Larger Living Spaces and Bedrooms:**
  - Strongly influence home prices.
  - Consider renovations to add bedrooms or expand living space for significant value increase.
- **Moderate Influences:**
- **Bathrooms and Lot Size:**
  - Have a moderate impact on home value.
  - Renovations in these areas could still add value, but may not be as impactful as expanding living space or adding bedrooms.

- **Minor Influences:**
- **Number of Floors and House Condition:**
  - Have weaker impacts on home prices.
  - Consideration needed, but not as strong predictors compared to other factors.
- **Model Limitations:**
- **Not Perfect Predictions:**
  - High prediction error suggests other factors affecting home prices not accounted for.
  - Always consider local factors and market conditions.
- **Continuous Improvement:**
- **Refine Model and Data:**
  - Improve predictions by incorporating more relevant data.
  - Continuous refinement to capture more variables for better predictions.

Conclusion

Conclusion:

- Understanding key renovation areas can significantly impact home values.
- Consideration of model limitations crucial for informed decision-making.

Next Step

- Continuously refine model with more relevant data.
- Stay updated on local factors.